

IDENTIFICATION OF DNA FEATURES AT THE TRANSITION REGION OF VARIOUS CHROMATIN STATES

*A major project report submitted in partial fulfilment of the
requirement for the degree of*

Master of Technology

In

Bioinformatics

Submitted by

Puneet Rawat

REG. NO: DTU/13/M.TECH/363

Delhi Technological University, Delhi, India

Under the supervision of

Dr. Monica Sharma



Department of Biotechnology
Delhi Technological University
(Formerly Delhi College of Engineering)
ShahbadDaultapur, Main Bawana Road, Delhi-110042, INDIA



CERTIFICATE

This is to certify that the M. Tech. dissertation entitled “**Identification of DNA features at the transition region of various chromatin states**”, submitted by **Puneet Rawat (DTU/13/M.TECH/363)** in partial fulfilment of the requirement for the award of the degree of Master of Technology, Delhi Technological University (Formerly Delhi College of Engineering, University of Delhi), is an authentic record of the candidate’s own work carried out at Centre for Cellular and Molecular Biology, Hyderabad under the guidance of Dr. Rakesh K Mishra.

The information and data enclosed in this dissertation is original and has not been submitted elsewhere for honouring of any other degree.

Date: 04/07/2015

Dr. Monica Sharma
Assistant Professor
Department of Bio-Technology
Delhi Technological University
(Formerly Delhi College of Engineering,
University of Delhi)

Dr. Bansi Das malhotra
Professor and HOD
Department of Bio-Technology
Delhi Technological University
(Formerly Delhi College of Engineering,
University of Delhi)



CERTIFICATE

This is to certify that the M. Tech. dissertation entitled “**Identification of DNA features at the transition region of various chromatin states**”, submitted by **Puneet Rawat (DTU/13/M.TECH/363)** in partial fulfilment of the requirement for the award of the degree of Master of Technology, Delhi Technological University (Formerly Delhi College of Engineering, University of Delhi), is an authentic record of the candidate’s own work carried out by him under my guidance at Centre for Cellular and Molecular Biology, Hyderabad.

The information and data enclosed in this dissertation is original and has not been submitted elsewhere for honouring of any other degree.

Date: 30/06/2015

Dr. Rakesh K Mishra
Senior Principal Scientist and Group Leader
Centre for Cellular and Molecular Biology
Uppal Road, Hyderabad 500 007
India.

DECLARATION

I, **Puneet Rawat(DTU/13/M.TECH/363)** declare that M. Tech. dissertation entitled “**Identification of DNA features at the transition region of various chromatin states**”, submitted in partial fulfilment of the requirement for the award of the degree of Master of Technology, Delhi Technological University (Formerly Delhi College of Engineering, University of Delhi), is an authentic record of my own work carried out under the guidance of **Dr. Rakesh k. Mishra**, Senior Principal Scientist and Group Leader, Centre for Cellular and Molecular Biology, Hyderabad

The information and data enclosed in this dissertation is original and has not been submitted elsewhere for honouring of any other degree.

Date: 04.07.2015

Name: Puneet Rawat

Place: New Delhi

Signature:

ACKNOWLEDGEMENT

I take this opportunity to express my deep sense of gratitude and indebtedness to my humble mentor, teacher and benefactor “Dr. Monica Sharma”. Whose help, encouragement and constant critics keep my moral high and expand my knowledge to a greater extent in the field of bioinformatics. Without her assistance and guidance, it would have been difficult for me to shape up this work.

Also I express my thanks to “Dr. Rakesh k. Mishra”, for his constant encouragement and valuable guidance during this project work. I am very thankful “A Srinivasan” for his constant guidance and to all my friends here for making moments spent here to be cherished lifetime. Last but not least I wish to express my devotional gratitude for my parents for their constant support and encouragement during the completion of my studies.

Date: 4/07/2015

Place: New Delhi

Puneet Rawat

2k13/BIO/13

M.Tech (Bioinformatics)

Delhi Technological University

CONTENTS

TOPIC	PAGE NO
<i>LIST OF FIGURES</i>	7
<i>LIST OF TABLES</i>	8
<i>LIST OF ABBREVIATIONS</i>	9
1. ABSTRACT	9
2. INTRODUCTION	11
3. REVIEW OF LITERATURE	15
3.1. ENCODE Project	
3.2. Histone modification	
3.3. ChIP-seq	
3.4. Computational analysis of ChIP-seq experiment	
3.5. Analysis of Peak calling algorithm	
4. METHODOLOGY	21
5. RESULT	28
6. DISCUSSION AND FUTURE PERSPECTIVE	30
7. REFERENCES	31
8. APPENDIX- 1(scripts)	34
9. APPENDIX- 2 (tables)	48

LIST OF FIGURES

Fig.1. DNA associates with histone proteins to form chromatin

Fig.2. *Drosophila melanogaster* genome

Fig.3. Pipeline showing the computational analysis of ChIP-seq experiment's results

Fig.4. ChIP-seq peak calling programs and comparison of their basic properties.

Fig.5. Comparative analysis of peaks identified.

Fig.6. Improper wig file format present in modENCODE dataset

Fig.7. Converted proper wiggle format

Fig.8. Flowchart of the peak finding algorithm

Fig.9. Different developmental stage dataset of drosophila compared with the peaks only:

- a. *Drosophila* embryo: whole chr2L chromosome compared with peaks only
- b. *Drosophila* larvae-L1: whole chr2L chromosome compared with peaks only
- c. *Drosophila* pupae: whole chr2L chromosome compared with peaks only
- d. *Drosophila* adult: whole chr2L chromosome compared with peaks only

Fig.10. Analyzing the results of the peak calling algorithm

LIST OF TABLES

Table 1: Histone modification in transcription regulation

Table 2: No. of peaks found for different datasets

Table 3: Motifs present in the peak regions probable cause of histone modification initiation

Table 4: Motifs present in the end regions, probable cause of histone modification termination

Table 5: Motifs present in the peak region of different dataset

Table 6: Motifs present in the end region of different dataset

LIST OF ABBREVIATIONS

- 1.** H3K27me3 : histone 3 lysine 27 trimethylation
- 2.** H3K4me3 : histone 3 lysine 4 trimethylation
- 3.** ChIP-seq : chromatin immuno-precipitation along with sequencing
- 4.** ENCODE : Encyclopaedia of DNA Elements
- 5.** modENCODE : model organism Encyclopaedia of DNA Elements
- 6.** .wig : wiggle format
- 7.** UCSC : university of California, santacruz
- 8.** GUI : graphical user interface
- 9.** NHGRI : National Human Genome Research Institute
- 10.** PRC : polycomb repressive complex
- 11.** CTD : C terminal domain
- 12.** TF : transcription factor
- 13.** MACS : model-based analysis for ChIP-seq
- 14.** GABP : growth-associated binding protein
- 15.** NRSF : neuron restrictive silencer factor
- 16.** FoxA1 : hepatocyte nuclear factor 3 α
- 17.** MEME : Multiple EM for Motif Elicitation
- 18.** DREME : Discriminative Regular Expression Motif Elicitation
- 19.** IUPAC : International Union of Pure and Applied Chemistry

IDENTIFICATION OF DNA FEATURES AT THE TRANSITION REGION OF VARIOUS CHROMATIN STATES

Puneet Rawat

Delhi Technological University, Delhi, India

ABSTRACT

Epigenetic factor plays an important role in regulation of gene expression and functions. ChIP-seq experiments (chromatin immune precipitation along with sequencing) are frequently used for mapping those epigenetic states throughout the genome. Histone modification is such an epigenetic factor which is analyzed using ChIP-seq experiment. Generally, data generated by these experiments are very noisy and diffused. Good signal values of ChIP-seq can range from few nucleotides to large domains. Histones are modified by post translational modifications such as methylation, acetylation, phosphorylation and ubiquitination, hence they influences the gene expression. One of modification H3K27me3 (histone 3 lysine 27 trimethylation) is known for shutting down the transcription when it tightly associates with inactive gene promoters by trimethylation. H3K27me3 data are very difficult to analyze compared to other datasets as they do not have very distinct signal peaks.

So by genome-level analysis of four different developmental stages of *Drosophila melanogaster* for H3k27me3 (ChIP-seq) data we are trying to identify the DNA signature that are responsible for the initiation and ending of those modifications. Our objective is the comparative study of motif enrichment within the signal regions and transition regions to find the possible motif for those histone modifications. H3K4me3 acts in opposition to H3K27me3 so it is used as a negative control.

INTRODUCTION

According to layman, **Epigenetics** is the study of environmental or external factors that turn genes *on* and *off* and affect how cells *read* genes.^[1] So, it can be said that epigenetics is the study of change in the transcription of gene within cell without changing the DNA sequence. Many epigenetic factors play an important role in regulation of gene expression like:

1. **Covalent modifications of DNA and histones:**^[2] DNA modifications (e.g. methylation of cytosine and hydroxymethylation) and histone proteins (e.g. acetylation of lysine, methylation of lysine and arginine, phosphorylation of serine and threonine, and lysine ubiquitination and sumoylation) are responsible for epigenetic inheritance. Chromatin is the complex of DNA and the histone proteins and different genes are regulated through its remodeling. Remodeling of chromatin is achieved through two main mechanisms:
 - i. **Post translational modification** of the histone protein. These changes in the amino acids might change the shape of the histone.
 - ii. **Addition of methyl groups** at CG rich regions (CpG islands), to convert cytosine to 5-methylcytosine. It is much like a normal cytosine. But, methylated areas will be less transcriptionally active.

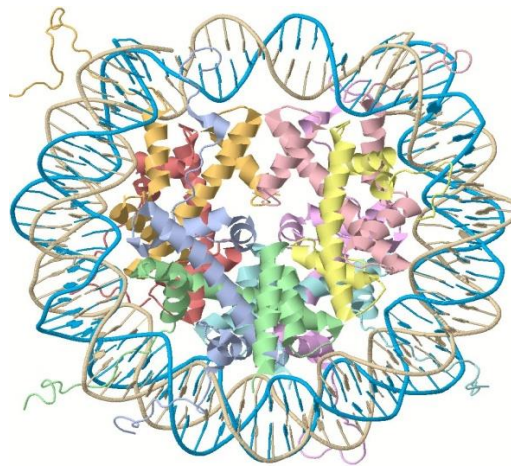


Fig.1. DNA associates with histone proteins to form chromatin.^[39]

2. **RNA transcripts and their encoded proteins:** Sometimes a gene turns on and transcribes a product that, through the transcription activity, maintains the activity of that gene^[3].
3. **MicroRNAs:** MicroRNAs (miRNAs) are members of small non-coding RNA (size: 17-25 nucleotide).^[4] Each miRNA expressed in a cell may down regulate about 100 to 200 messenger RNAs.^[5] Most of the downregulation of mRNAs is caused by decaying the

targeted mRNA, while some other occurs at the time of translation into protein.^[6]In humans, about 60% of the protein coding genes are regulated by miRNAs^[7].

4. **Prions:** these are infectious forms of proteins. Generally, proteins fold into specific shape to perform specific cellular functions, but some can form infectious conformation known as a prion. They have the ability of catalyst to convert other native state of that protein to an infectious conformational state. So they can be viewed as epigenetic agents that can change phenotype without modifying the genome^[8].
5. **Nucleosome positioning:** Eukaryotic DNA is packed and ordered with the help of protein coat, this structural unit is called nucleosomes. Their positions are not random; they determine the availability of DNA to regulatory proteins. This determines the level of gene expression and cell differentiation^[9].

modENCODE is the part of the **ENCODE** project which is working on identification of functional elements (epigenetic elements discussed above) in selected model organism genomes, specifically, *Drosophila melanogaster* and *Caenorhabditis elegans*. modENCODE contains the raw and interpreted data of different experiments. Histone modification H3K27me3 is related to the repression of the gene. ChIP-seq experiment data of H3K27me3 modification in *Drosophila melanogaster* is analyzed. H3K4me3 modification works opposite of the H3K27me3 modification. So it is used as negative control for comparative studies.

DROSOPHILA GENOME:

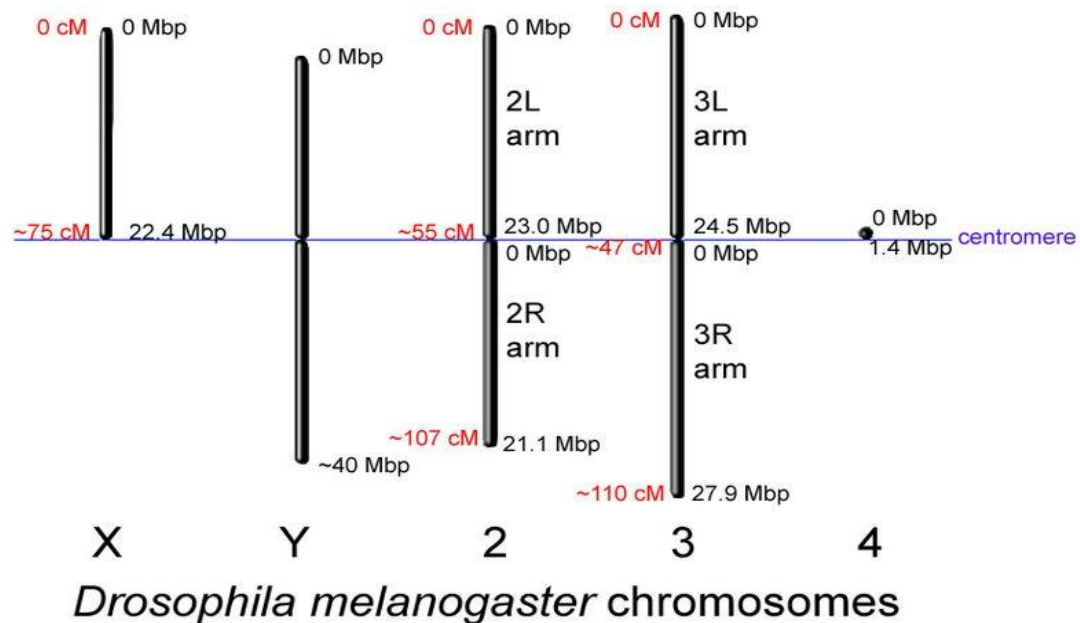


Fig.2. *Drosophila melanogaster* genome^[40]

Drosophila melanogaster is a fly species taxonomic order Diptera and family Drosophilidae. Generally, it is known as common fruit fly or vinegar fly. Charles W. Woodworth proposed to use this species as a model organism, since then it is used widely for biological research in the field of physiology, genetics, microbial pathogenesis, and biological evolution. The reasons to use this as model organism are: breeds quickly and lays many eggs, easy to care, have four pairs of chromosomes.^[10]

D. melanogaster genome was sequenced in 2000, and was curated at FlyBase database^[11]. It contains four pairs of chromosomes: an X/Y sex chromosome pair and other autosomes labeled 2, 3, and 4. The last chromosome is tiny chromosome, so it is often ignored. The *D. melanogaster*'s 139.5 million base pair genome has been annotated^[12] and according to Ensemble release 73, it contains around 15,682 genes. More than 60% of the genome of *Drosophila* appears to be functional but non-protein-coding DNA^[13] which is involved only in gene expression control. Human disease genes have a recognizable match of about 61% with the genome of fruit flies.^[14] Around 50% of fly protein sequences have homologs in mammals.

UCSC Genome Bioinformatics:

This website provides graphical user interface for visualization and database access for variety of vertebrate and invertebrate species and major model organisms. UCSC Genome Browser is used to visualize the genomic or custom tracks in the GUI form using the chromosome coordinates while Table Browser gives us convenient access to the underlying databases and information. We are using wiggle file format for the analysis purpose.

Wiggle file format:

It is a line oriented format starts with "track type=wiggle_0", which identifies the track as a wiggle track. In this track definition line we can add many options to control the default display.

Second line of the wiggle format is declaration lines and others are data lines. There are two ways to format wiggle data: variableStep and fixedStep. These formats are designed to provide convenient and compact way to write the file:

- **variableStep** is for data with irregular intervals between each of the data entry. This is most commonly used wiggle format. After the track definition line of wiggle format, variable step begins with a declaration line which contains "variableStep" at first followed by chromosome name and span value. Data lines are followed by two columns containing positions of chromosome and signal values. The span value begins at each

chromosome's position and indicates the number of bases (default span: 1) that data value should cover.

For example:

```
variableStep[\t]chrom=chrN[\t][span=windowSize][\n]  
chromStartA[\t]dataValueA[\n]  
chromStartB[\t]dataValueB[\n]
```

- **fixedStep** is for data which has regular intervals between chromosome positions. These are more compact wiggle format. After track definition line, fixed step begins with a declaration line which contains “fixedStep” itself and then start coordinate, chromosome name and span value. Then it is followed by a single column of signal data values. The span argument has the same meaning as in variable step format.

For example:

```
fixedStep[\t]chrom=chrN[\t]start=position[\t]step=stepInterval[\t][span=windowSize][\n]  
dataValue1[\n]  
dataValue2[\n]
```

REVIEW OF LITRETURE

3.1 ENCODE PROJECT:

The **Encyclopedia of DNA Elements (ENCODE)** project is launched by the National Human Genome Research Institute (NHGRI), US in September 2003 ^[15]. This public research project was started as a continuation to the Human Genome Project. This project targets to understand different functional elements or epigenetic element in the human genome. The project is a consortium of the research groups from all over the world. Any person can access the data generated from this project through public databases.

Humans have approximately 20,000 protein-coding genes, which are only about one and half percent of the DNA in the human genome. ENCODE project primarily focus on, “what is the role of the other component of the genome”, which was traditionally regarded as "junk DNA". There are many reasons to target those regions like: approximately 90% of SNP (single-nucleotide polymorphism) in the human genome that have already been linked to various diseases, are found outside of protein-coding regions.^[16]

Protein-coding gene's activity can be controlled by promoter, transcriptional regulatory sequences, chromatin structure regions and histone modification. Identifying the location of these regulatory elements and their effect on gene transcription could find relation between gene expression variation and disease development.^[17]

modENCODE project^[18]: The Model Organism ENCYclopedia Of DNA Elements (modENCODE) is extension of the original ENCODE project which targets the identification of epigenetic elements in specific model organisms like *Drosophila melanogaster* and *Caenorhabditis elegans*. This extension of ENCODE project permits validation of findings of the ENCODE project, that is very difficult to do in humans. modENCODE contains raw and interpreted data from different experiments. At the moment, modENCODERun as a Research Network formed by 11 primary projects. They are divided between worms and flies and it spans the following:

1. Gene structure
2. mRNA and ncRNA expression profiling
3. Transcription factor binding sites
4. Histone modifications and replacement
5. Chromatin structure
6. DNA replication initiation and timing
7. Copy number variation.^[19]

3.2 Histone modification:

Histones are alkaline proteins that are found in eukaryotic cell nuclei. Their packed and ordered structural form with DNA is called nucleosome.^[20] Histones are the main component of chromatin, winds around the DNA, and play important role in gene regulation.

Posttranslational modifications alter interaction of histones with DNA. Covalent modifications on long tails of H3 and H4 include methylation, phosphorylation, acetylation, ubiquitination, SUMOylation, ADP-ribosylation and citrullination (deimination). The core of the histones H3, H2B and H2A can be modified.^[21] Histone modifications act in regulation of genes, repair of DNA, condensation of chromosomes (mitosis) and spermatogenesis (meiosis).^[22]

Nomenclature of the histone modification:

- The histone name (e.g., H2)
- The single-letter abbreviated form of amino acid (e.g., K for Lysine) and the position of amino acid in the protein (e.g. 4,27)
- The modification type (P: phosphate, Ub: ubiquitin, Me: methyl, Ac: acetyl)
- The number of modifications per residue (only Me is known to occur more than once. 1, 2 or 3 is mono-, di- or tri-methylation)

Type of modification	Histone							
	H3K4	H3K9	H3K14	H3K27	H3K79	H3K36	H4K20	H2BK5
mono-methylation	activation	activation		activation	activation		activation	activation
di-methylation		repression		repression	activation			
tri-methylation	activation	repression		repression	Activation, repression	activation		repression
acetylation		activation	activation	activation				

Table.1. Histone modification in transcription regulation^[41]

Trimethylation of H3 lysine 27 (H3K27Me3): H3K27 modification of histone is deposited by the PRC2 complex. It is a gene repression marker, and is likely to exert a repressive function, PRC1 is another polycomb complex that can bind H3K27me3^[23]. It is recruited through the action of PRC2 and adds in the modification of histone H2AK119Ub which aids compaction of chromatin.^[24]

Trimethylation of H3 lysine 4 (H3K4Me3)^[25]: COMPASS complex performs the H3K4 trimethylation.^[26] It is a marker of active promoter and the extent of this modification at a gene's

promoter relates with transcriptional activity of the gene. The relation of this mark with transcription is rather convoluted, early in gene transcription, RNA polymerase II shifts from ‘initiating’ to ‘elongating’, which is marked by phosphorylation state change of the RNA polymerase II CTD (C terminal domain). Enzyme that phosphorylates the RNA polymerase II CTD also phosphorylates the Rad6 complex,^[27] which adds a mark to H2B K123Ub (in mammals: K120).^[28] COMPASS requires this mark to trimethylate H3K4 at promoters.^[25]

3.3 ChIP-seq:

ChIP-sequencing, also known as ChIP-Seq, is an experimental technique used to analyze protein-DNA interactions. ChIP-seq combines chromatin immune-precipitation (ChIP) with parallel DNA sequencing to mark the DNA-protein complex sites. It can precisely map sites of protein on genome wide scale. ChIP-seq is used to identify how epigenetic factors like histone protein and transcription factors influence expression of gene.

Workflow of ChIP-seq:

ChIP is a method to enrich only those DNA sequences which are bound by a specific protein in living cells. The ChIP process enriches specific cross-linked DNA-protein complexes using an antibody against that protein. Short nucleotide adaptors are then added to the small length DNA that were bound to our protein of interest to enable parallel sequencing.

Seq (sequencing): After selecting the size, all the DNA fragments from the ChIP experiment are sequenced using a genome sequencer. Whole genome can be sequenced in a single run with high resolution, to locate feature precisely on the chromosomes. There are many new sequencing techniques available. Such as: on a solid flow cell substrate, amplifying the cluster of adapter-ligated ChIP DNA fragments

3.4 Computational analysis of ChIP-seq experiment^[29]:

During the sequencing step, raw nucleotide short tags (aka reads) are sequenced and digitalized base-by-base in the ChIP-seq. Moreover In ChIP-seq experiment linker/adaptor contamination, background noise and image processing error, all contribute to the ChIP-seq error profile.

Single-ended short reads of size 25–35 bps^(Fig.3.step-1) are commonly used in ChIP-seq studies. We define a mappable read as those which aligns to a unique location in the genome; non-repeating ones are the mappable reads that occur only once in the dataset. In ChIP-seq experiment, goal is to gain an adequate number of mappable reads. At current Solexa/Illumina capacity, a single sequencing lane yields tens of millions of short reads and approximately half of them can be uniquely aligned back to the reference genome. In mammalian genomes, coverage of 10 million reads typically provides clear binding signals at a large fraction of the binding sites. For smaller genome such as *Drosophila melanogaster*, one can attain higher signal intensity at the same sequencing depth. Before any large scale production run of a ChIP-seq experiment, it is useful to first conduct a demo or pilot experiment run on a single lane to confirm the experiment output and to save the valuable time and resource.

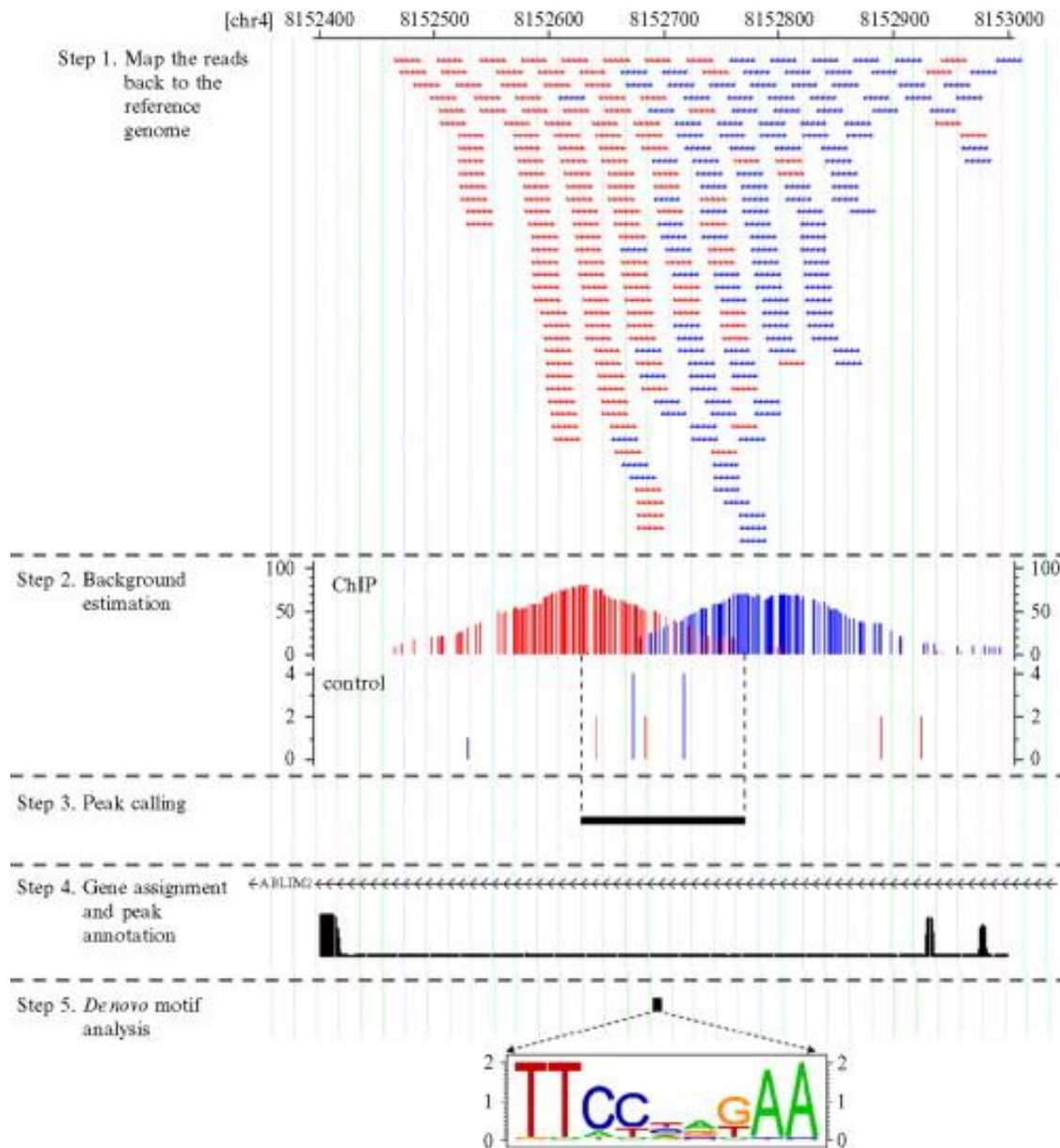


Fig.3. pipeline showing the computational analysis of ChIP-seq experiment's results^[29]

Peak calling^[29]:

This step is to identify the ChIP signal enriched genomic regions. In a successful experiment, if a protein factor has a sharply focused binding site, one should be able to observe the bi-horned bell-shaped peaks. These peaks will be shaped at both the Watson strand (represents the 5'-end of a ChIP fragment and in the left red color) and the Crick strand (represents the 3'-end and in the right blue color) (Fig.3.step-2). Thus the two peaks may be used to define a candidate binding region.

As depicted in Fig. 3.1, to obtain the signal profile, we use a fixed window size w and count the number of the Watson and the Crick reads that fall into each nonoverlapping window along the entire genome. Window size $w = 100$ is recommended for sequence-specific TF-binding ChIP-

seq data [30]. The mapped reads and the signal profile (defined below) can be visualized in a genome browser as UCSC genome browser. [31]

De novo motif analysis[29]:

Another important task in the analysis of the predicted peak regions is de novo motif discovery. In some studies, the exact sequence to which the TF binds is known, or a set of validated binding sites is available. If this information is not available, we can extract the sequences from the peak regions.

3.5 Analysis of Peak calling algorithm:

There are many peak calling algorithms present in both professional and open source domain. In an article different peak calling algorithms are compared for their performance in a dataset [32]. For given datasets, each algorithm had different stringency and no. of peaks.

Program	Version	Graphical user interface?	Window-based scan	Tag clustering	Gaussian kernel estimator	Strand-specific density	Peak height or fold enrichment (FE)	Background subtraction	Compensates for genomic duplications or deletions	False Discovery Rate	Compare to normalized control data (FE)	Compare to statistical model fitted with control data	Statistical model or test
CisGenome	1.1	X*	X			X	X		X		X		conditional binomial model
Minimal ChipSeq Peak Finder	2.0.1		X			X				X			
E-RANGE	3.1		X			X				X	X		chromosome scale Poisson dist.
MACS	1.3.5		X			X			X		X		local Poisson dist.
QuEST	2.3			X		X			X**		X		chromosome scale Poisson dist.
HPeak	1.1		X			X					X		Hidden Markov Model
Sole-Search	1	X	X			X		X			X		One sample t-test
PeakSeq	1.01		X			X					X		conditional binomial model
SISSRS	1.4		X		X					X			
spp package (wtd & mtc)	1.7		X		X		X	X'	X				
			Generating density profiles		Peak assignment		Adjustments w. control data		Significance relative to control data				

X* = Windows-only GUI or cross-platform command line interface
X** = optional if sufficient data is available to split control data
X' = method excludes putative duplicated regions, no treatment of deletions

Fig.4. ChIP-seq peak calling programs and comparison of their basic properties.[32]

There were 3 different dataset GABP, FoxA1 and NRSF. For each dataset results were obtained using algorithm with default or recommend settings. Core peaks in the Fig.5 are the consensus of all the peaks of different algorithm. True peaks were also identified by the qPCR experiment and the results were compared.

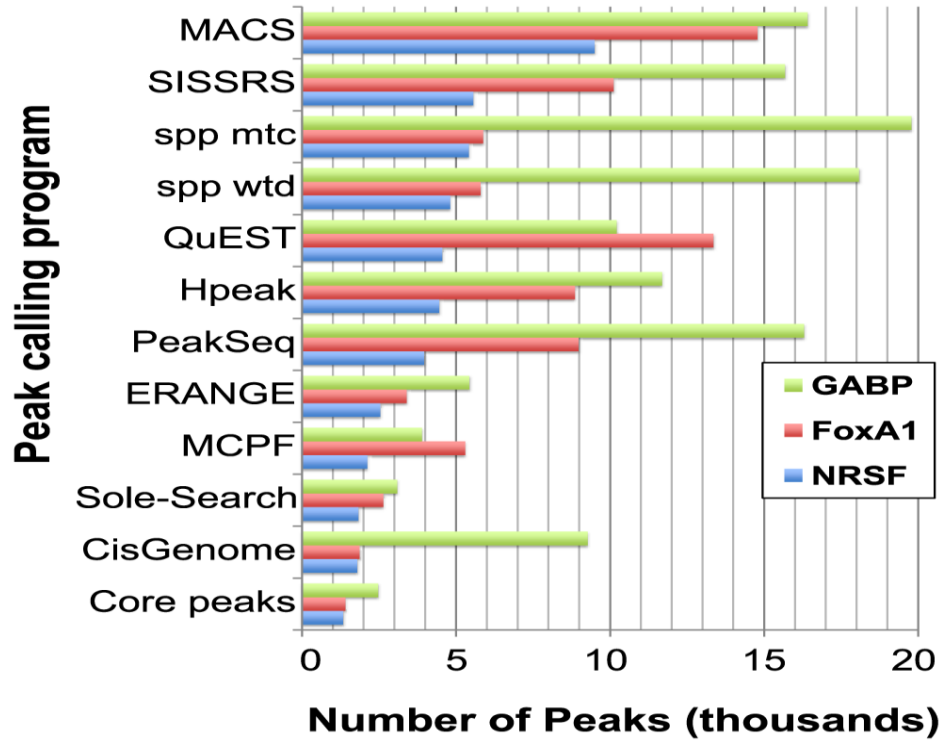


Fig.5 Comparative analysis of peaks identified.^[32]

When those peaks were compared with the qPCR results (positive peaks), it was found that for different dataset they have varying results. Some algorithm could not identify few positive peaks also like in NRSF dataset; Sole-Search and CisGenome had less sensitivity. MACS was found with the best result for given dataset. By analyzing these result, need of a customized algorithm is deeply felt because of the low sensitivity of those algorithms. We are also interested in the false peaks as they have signal enrichment. The sequence signature there must have caused the increase in the signal intensity and more signatures will give us better results.

MATERIAL AND METHODOLOGY

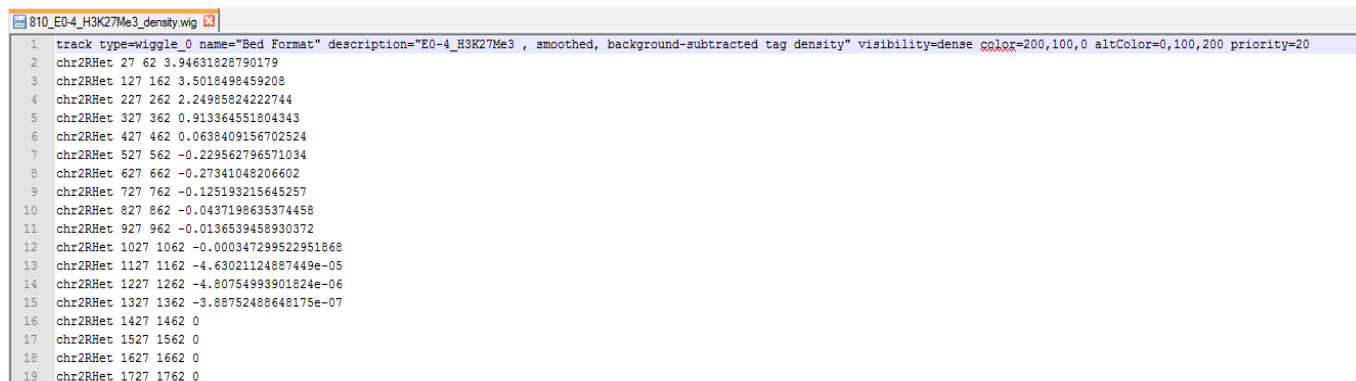
ModENCODE database generally contains the ChIP-seq experimental data in 3 different categories:

- **Raw data files:** these data files contain the unprocessed data directly from the experiments. Format generally contains the short segment of the sequences and their respective enrichment values.
e.g. *.sam, *.fastqetc
- **Signal data files:** from the raw data signal files are generated using some algorithm. Instead of the sequences, it contains the coordinates of the sequences in the genome and its corresponding signal values.
e.g. *.wig, *.bed etc
- **Interpreted data files:** these contain the valuable information interpreted from the signal or raw data files such as peaks coordinates etc.
e.g. *.gff etc

We are using signal data files for peak finding because they are less computation expensive. Dataset for the H3K27me3 and H3K4me3 modification were downloaded from the dataset (<http://data.modencode.org/>) in ModENCODE. We have taken four different developmental stages of Drosophila for the H3K27me3 data analysis:

- i. Embryo (0-4 hrs)
- ii. Larvae (L1 phase)
- iii. Pupae
- iv. Adult (male)

And as a control H3K4me3 histone modification dataset were taken for Adult male fly. Signal data files were in improper .wig format. The data inside the .wig files was actually formatted in .bed format. Following is the original .wig format:



```
810_E0-4_H3K27Me3_density.wig
1 track type=wiggle_0 name="Bed Format" description="E0-4_H3K27Me3 , smoothed, background-subtracted tag density" visibility=dense color=200,100,0 altColor=0,100,200 priority=20
2 chr2RHet 27 62 3.94631828790179
3 chr2RHet 127 162 3.5018498459208
4 chr2RHet 227 262 2.24985824222744
5 chr2RHet 327 362 0.913364551804343
6 chr2RHet 427 462 0.0638409156702524
7 chr2RHet 527 562 -0.229562796571034
8 chr2RHet 627 662 -0.27341048206602
9 chr2RHet 727 762 -0.125193215645257
10 chr2RHet 827 862 -0.0437198635374458
11 chr2RHet 927 962 -0.0136539458930372
12 chr2RHet 1027 1062 -0.000347299522951868
13 chr2RHet 1127 1162 -4.63021124887449e-05
14 chr2RHet 1227 1262 -4.80754993901824e-06
15 chr2RHet 1327 1362 -3.88752488646175e-07
16 chr2RHet 1427 1462 0
17 chr2RHet 1527 1562 0
18 chr2RHet 1627 1662 0
19 chr2RHet 1727 1762 0
```

Fig.6. improper wig file format present in modENCODE dataset

So our first task was to convert this in proper wiggle format. So we wrote a perl script “**bed_to_wig_convertor.pl**”(appendix-1). This has given us proper wig format:

```
proper_wig_format.wig
1 track type=wiggle_0 name="Bed Format" description="E0-4_H3K27Me3 , smoothed, background-subtracted tag density" visibility=dense color=200,100,0 altColor=0,100,200 priority=20
2
3 # this is a wiggle file format not a bed file format
4 variableStep chrom=chr2RHet span=100
5 27 3.94631828790179
6 127 3.5018498459208
7 227 2.2498582422744
8 327 0.913364551804343
9 427 0.0638409156702524
10 527 -0.229562796571034
11 627 -0.27341048206602
12 727 -0.125193215645257
13 827 -0.0437198635374458
14 927 -0.0136539458930372
15 1027 -0.000347299522951868
16 1127 -4.63021124887449e-05
17 1227 -4.80754993901824e-06
18 1327 -3.88752488648175e-07
19 1427 0
```

Fig.7.converted proper wiggle format

This wig file is then divided into separate files on the basis of chromosomes using “**file_division.pl**” (appendix-1). Then we analyzed the peaks for each chromosome. There are many reasons for developing our own peak finding algorithm. Peak calling algorithms currently present sometimes misses the important information due to higher stringency [32]. Almost every peak calling algorithm could not identify the actual length of the peak. They chop out the smaller values at the end of the peaks. For us, false peaks have as much importance as positive peaks as we are interested in the signatures that are causing the peaks. Generally peak calling algorithm does not consider these less value peaks. Yet we also have to keep it in mind that background noises should not be considered as a peak.

Peak calling algorithm:

Peak calling perl script “**peak.pl**”(appendix-1) should be used separately for each chromosome. This script is automated but it can take three arguments with the file name to provide flexibility in stringency:

- a. **-s:** this argument will define the minimum percentage of the total data point that will be included in the peak region. Basically this will define the stringency in the dataset. Lesser the value of this argument will lead to higher stringency. By default only the 20% (.2) of the total data is allowed in the peak regions. Data given should be in range of .01 to .99.
- b. **-p:** minimum peak length. By default the value is 5. The minimum value of the peak length should be greater than 5.
- c. **-h:** minimum peak height. This will define, what should be the minimum height of the peak compared to the height of the maximum peak. By default it is set 30% of the height of the maximum peak. Input value should be in range of .01 to .99

Example:test_file.wig -s0.4 -p10 -h0.6

Peak finding algorithm's flowchart:

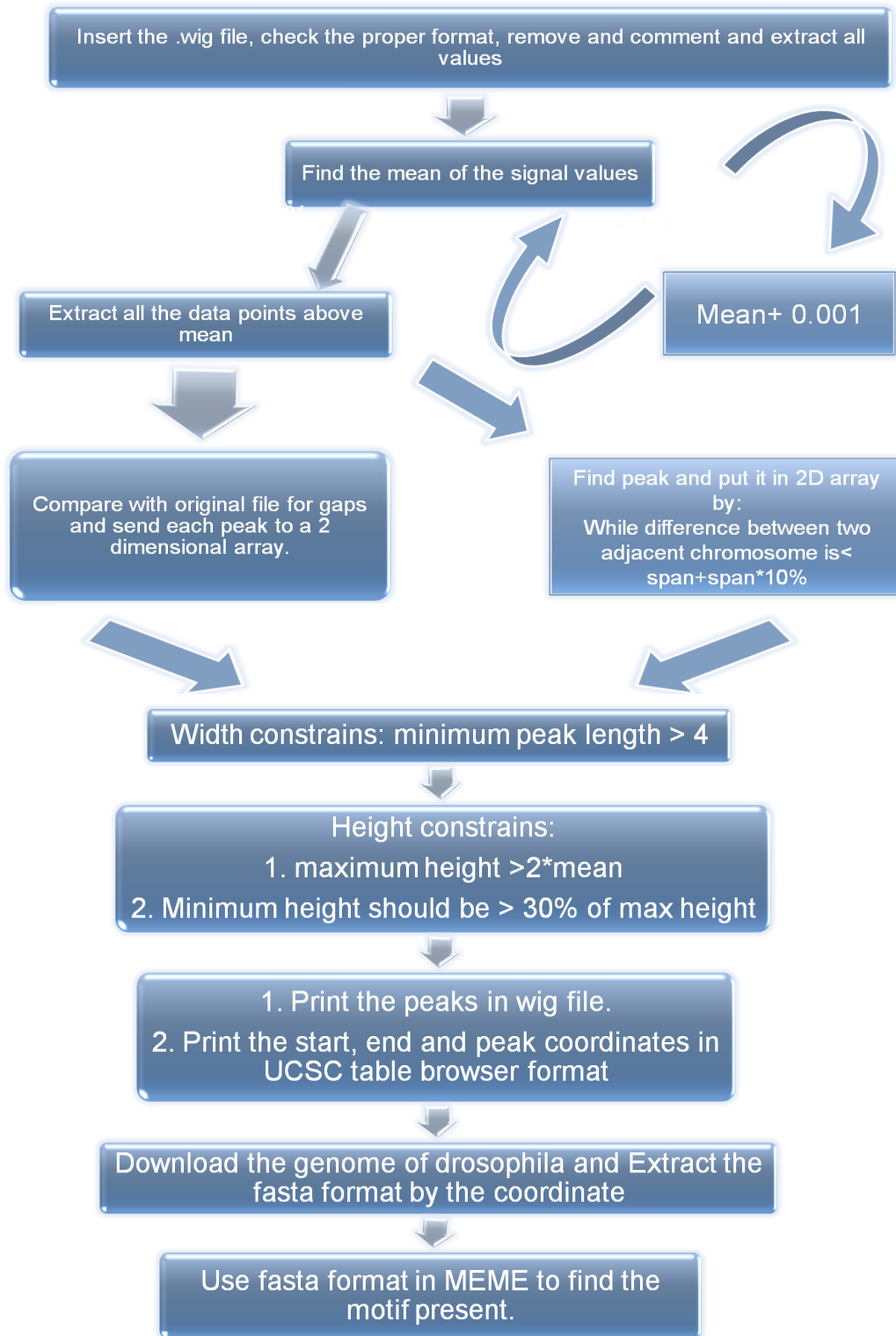


Fig.8. flowchart of the peak finding algorithm

We will run this for each chromosome. The results of each chromosome will automatically append on the previous results. It is better to remove any previous result files before using the algorithm for the first time. This script also produces a log file for each chromosome which contains the basic information about the input file and parameters used in the algorithm. This will be displayed on the screen also.

Limitation of this algorithm is that it is not suitable for the too much negative data because script does not take negative mean value. So in case of negative mean it changes mean to 0.1. If the mean is exactly 0.1 then try to reduce the value of `-s` parameter less than the default value (.2), so eventually mean value increase more than 0.1. But it will also increase the stringency and hence the upstream and downstream regions will not be that perfect. To overcome this problem we have already taken much larger region so we won't miss important motifs.

Visualization of the peaks:

UCSC Genome Browser:

The UCSC Genome Browser is an on-line tool to visualize the genome. It is hosted by the University of California, Santa Cruz (UCSC).^[33] It is an interactive web tool which offers access to data of genome sequence from a variety of species and major model organisms, it is also integrated with a large collection of aligned annotations with genome sequences. This tool is a GUI based open source web-based tool built on MySQL database for rapid visualization and querying.

UCSC Genome Browser is used to visualize the genomic or custom tracks in the GUI form through the chromosome coordinates while Table Browser gives us convenient access to the underlying databases and information.

After running the scripts we analyzed the results in UCSC genome browser as a custom track.



Fig.9.a.Drosophila embryo: whole chr2L chromosome compared with peaks only

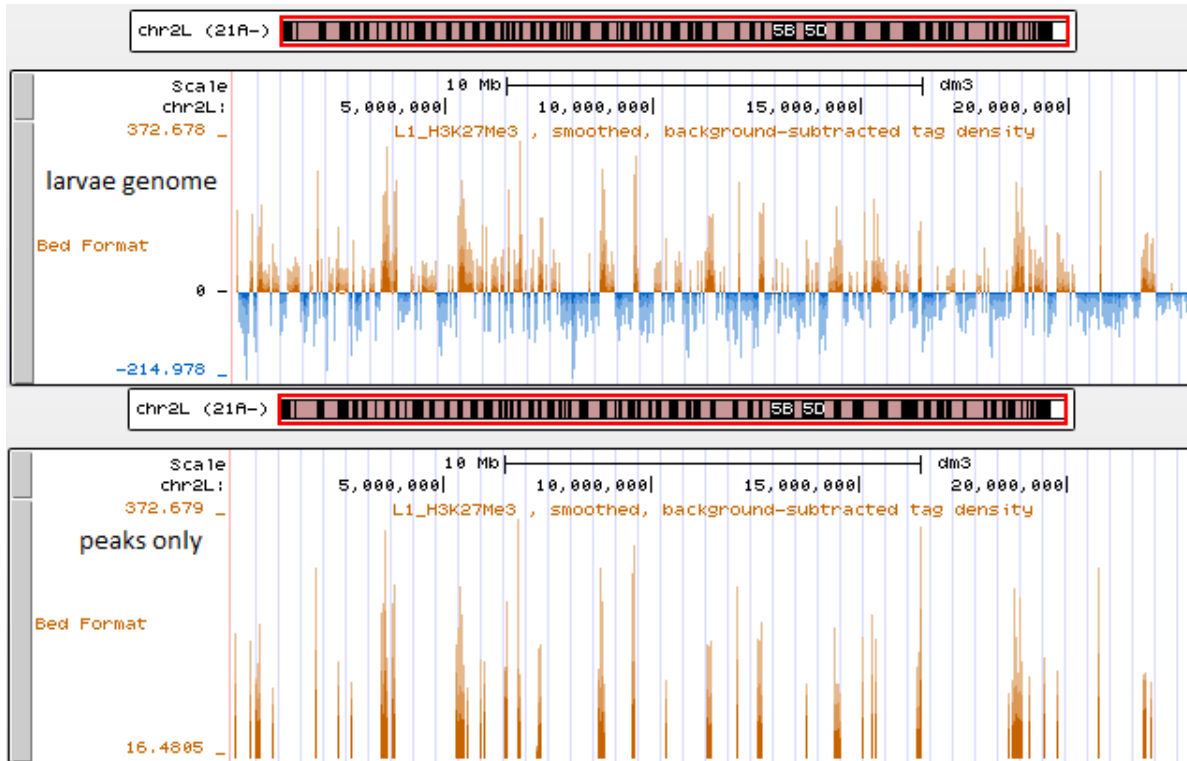


Fig.9.b. Drosophila larvae-L1: whole chr2L chromosome compared with peaks only

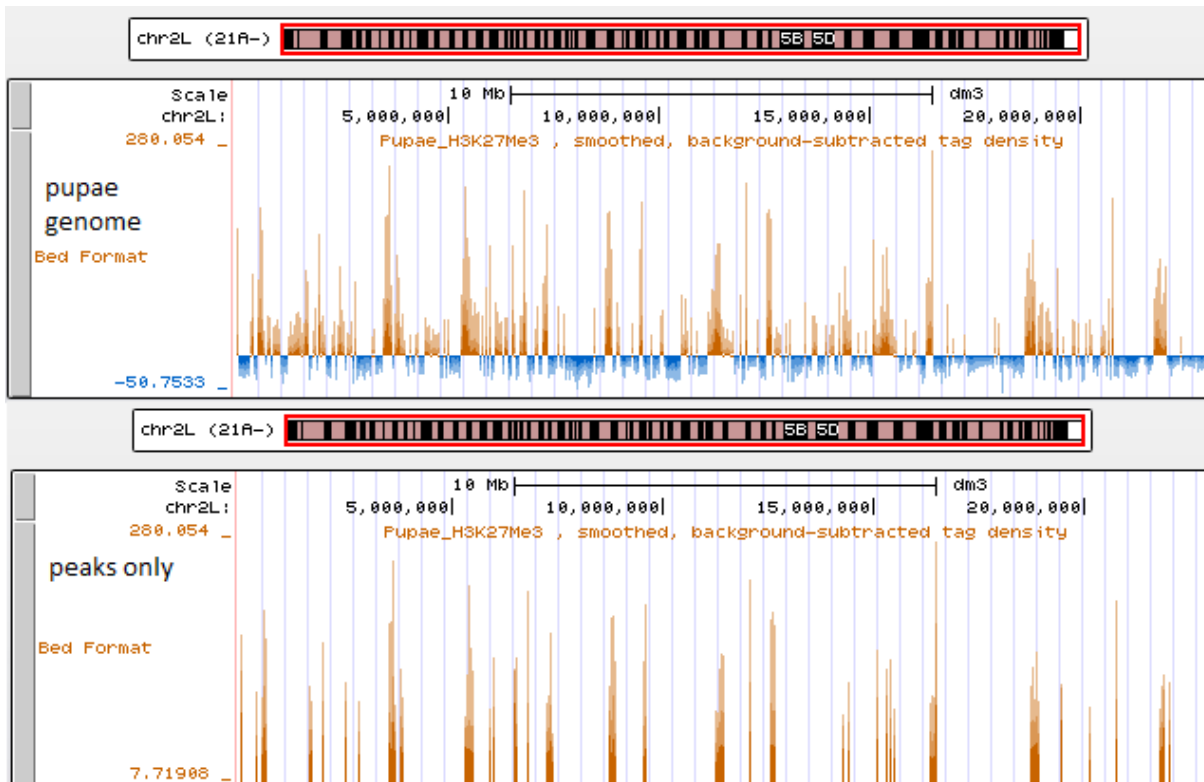


Fig.9.c. Drosophila pupae: whole chr2L chromosome compared with peaks only

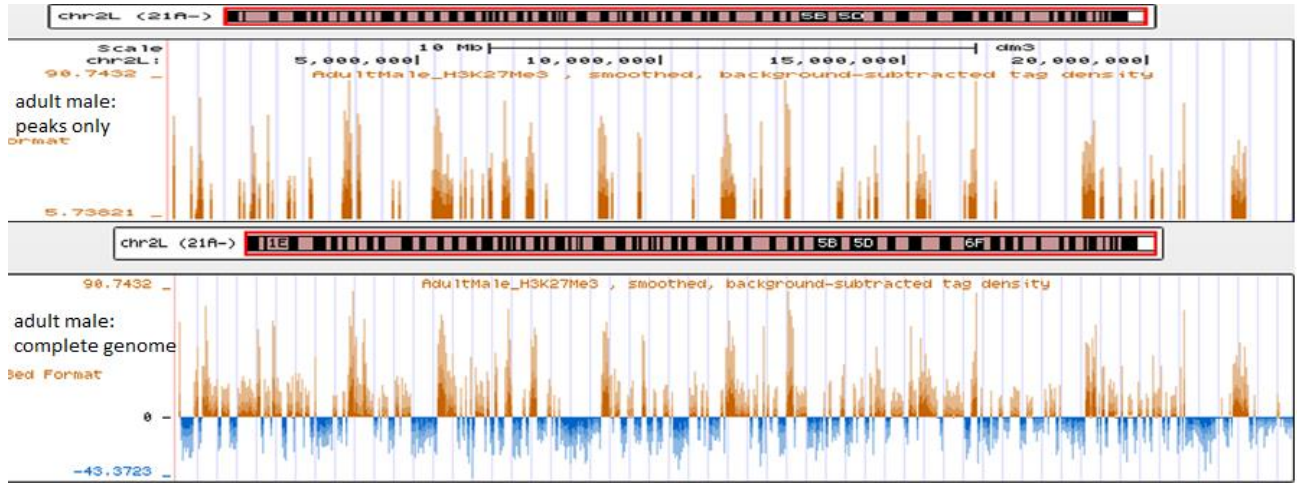


Fig.9.d. Drosophila adult: whole chr2L chromosome compared with peaks only

In depth analysis of larvae



Coordinates in chromosome:
chr2L:850,018-958,371
A histone modification site containing a gene.
(Best case scenario)

Coordinates in chromosome:
chr2L:18,756,855-18,843,853
An imperfect histone modification site
(worst case scenario)

Fig.10. Analyzing the results of the peak calling algorithm

Extracting the region of interest:

In the script “**peaks.pl**”, we have extracted the co-ordinates of the region of interest. (500 base pair upstream from starting point of peak and 500 base pair downstream from ending of peak) the result file is formatted in the UCSC Table Browser format.

Next step is to download the genome of *Drosophila melanogaster*. It was downloaded from the ftp server of **FlyBase database release 5.32** (same as used in preparing the modENCODE dataset).

From the genome, short sequences were taken out in **fasta format** on the basis of chromosome’s co-ordinates using the “**fastaconversion.pl**”^(appendix-1)script. Now we have the regions of interest and now we will try to identify some motif present on those regions.

Motif Analysis:

MEME-DREME:Multiple EM for Motif Elicitation or MEME is a tool for identification of motifs in a group of related DNA.^[34]A motif is a sequence pattern that occurs repeatedly in a group of related protein or DNA sequences. DREME is the part of MEME suite for motif identification in large data.

DREME or Discriminative Regular Expression Motif Elicitation ^[35]is a tool generally used for identification of motifs in large collections of sequences. It is very computationally efficient and therefore is best for motif search on large data sets taken from ChIP-seq experiments. To make it computationally efficient, DREME finds motifs which can be expressed only in the IUPAC alphabet, that means it contains the standard DNA alphabet ATGC as well as eleven combination characters (for example, Y indicates either C or T).

DREME discovers short, ungapped motifs that are relatively enriched in your nucleotide sequences compared with shuffled sequences or your control sequences. We provide 2 input files to the DREME tool:

- 1) Fasta file (.fa format) containing the peak regions: to identify the motifs causing the histone modification.
- 2) Fasta file (.fa format) containing the end regions of the peak: (we merge the start regions and end regions of the peaks into a single file): to identify the motifs stopping the histone modification.

We have used default E-value threshold of 0.05 and entries up to 50 motifs.

RESULTS

Peak Calling Algorithm:using peak calling algorithm on the given dataset produces 3 results files:

1. Wiggle file containing all data point of peaks. This file can be visualized in UCSC Genome Browser to check the results
2. Peak region and end region chromosome coordinates in UCSC Table Browser format. So the sequence can be extracted from the UCSC table browser or *Drosophila melanogaster* genome.
3. Log file to see the different parameter and properties used.

modENCODE Dataset	Developmental Stage	No of Peaks
modEncode_810	Embryo(0-4 hours)	484
modEncode_816	Larvae L1	1207
modEncode_819	Pupae	826
modEncode_820	Adult Male	1622
modEncode_800 (H3k4Me3 modification site- for comparison/control)	Adult Male	2570

Table.2. No. of peaks found for different datasets

Motif analysis:

We got the sequences in fasta format from the chromosome coordinates. Now we used DREME online tool of MEME SUITE. We have used default e-value threshold of 0.05 and limited maximum motif up to 50. We looked for common motifs in H3K27me3 dataset in different developmental stage and extracted those motifs which are not present in the control. Motifs present in the peak regions are most probably responsible for initiation of H3K27me3 histone

modification while the end region motifs might be responsible for stopping the histone modification.

Motif found in the peak regions



Table.3. Motifs present in the peak regions probable cause of histone modification initiation

Few of above motifs are experimentally proved such as: CCC(C/G)CCCC, CTCCTCC⁽³⁶⁾and CACACAC(A/T/C)⁽³⁷⁾.

Motif found in the end regions

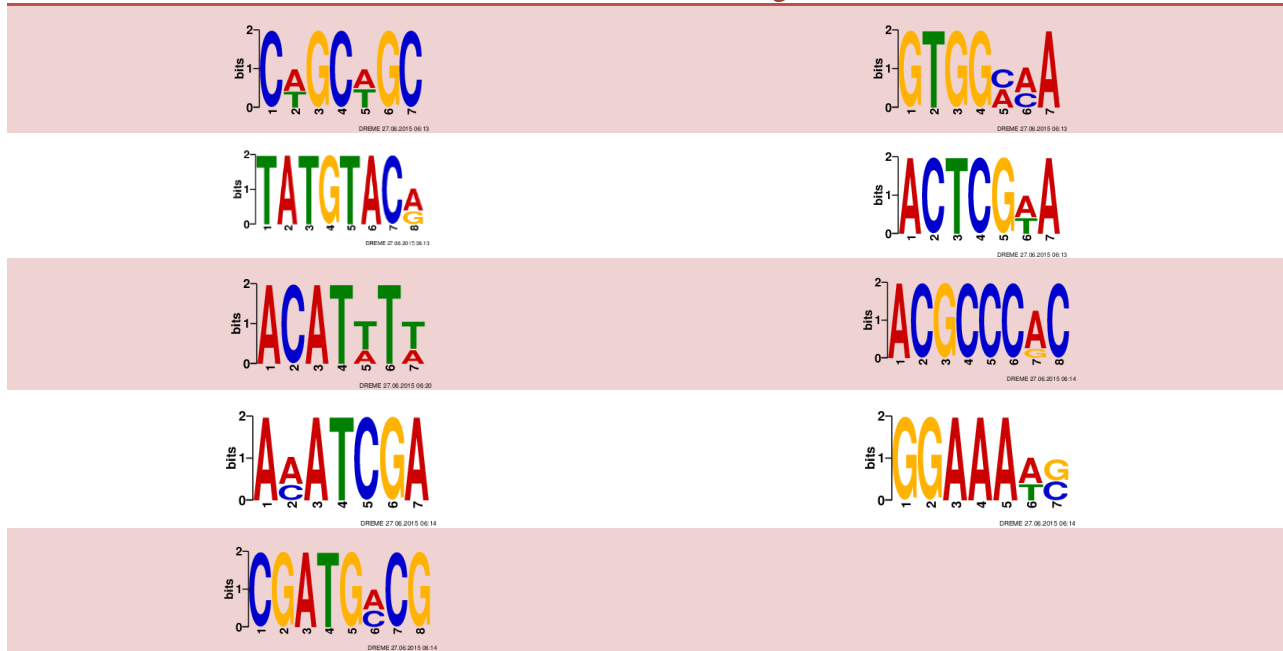


Table.4. Motifs present in the end regions,probable cause of histone modification termination

DISCUSSION AND FUTURE PERSPECTIVE

We aimed to find the motifs or the DNA features that are enriched at the transition regions defined by various chromatin states. For this we used the H3k27me3 histone modification data sets from modENCODE.

We have identified several motifs that are enriched at the transition region. We also looked for motifs that are over represented within the peak region. The analyzed data sets include various developmental stage of *Drosophila*. We have identified some true motif enriched in their respective regions but the rate of the false motif discovery is high in the results. It is mostly because of the false mapping of peak borders or transition region, and splitting of larger peaks may be the reason behind this.

Identifying the peaks in histone modification data of H3K27 is a difficult task compared to other histone modification due to false peak, noisy background and large domains and experimental inefficiency. With this program we are able to partly achieve that. However, further improvements are required to find the broader peaks and their correct transition regions. Upon improvement, this program can be applied on any other histone modification data sets and their associated motifs could be identified afterwards.

REFERENCES

1. "Epigenetics". Icahn School of Medicine at Mount Sinai. Retrieved 26 May 2015.
2. Ptashne M (April 2007). "On the use of the word 'epigenetic'". *Curr. Biol.* 17 (7): R233–6. doi:10.1016/j.cub.2007.02.030. PMID 17407749.
3. Mattick JS, Amaral PP, Dinger ME, Mercer TR, Mehler MF (January 2009). "RNA regulation of epigenetic processes". *BioEssays* 31 (1): 51–9. doi:10.1002/bies.080099. PMID 19154003.
4. Bernal JE, Duran C, Papiha SS (2012). "Transcriptional and epigenetic regulation of human microRNAs". *Cancer Lett* 331 (1): 1–10. doi:10.1016/j.canlet.2012.12.006. PMID 3246373.
5. Lim LP, Lau NC, Garrett-Engle P, Grimson A, Schelter JM, Castle J et al. (2005). "Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs". *Nature* 433 (7027): 769–773. doi:10.1038/nature03315. PMID 15685193.
6. Lee D, Shin C (2012). "MicroRNA-target interactions: new insights from genome-wide approaches" *Ann N Y AcadSci* 1271:118-28. doi: 10.1111/j.1749-6632.2012.06745.x. Review. PMID 23050973
7. Friedman RC, Farh KK, Burge CB, Bartel DP (2009). "Most mammalian mRNAs are conserved targets of microRNAs". *Genome Res* 19 (1): 92–105. doi:10.1101/gr.082701.108. PMC 2612969. PMID 18955434.
8. Yool A, Edmunds WJ (1998). "Epigenetic inheritance and prions". *Journal of Evolutionary Biology* 11 (2): 241–242. doi:10.1007/s000360050085
9. Teif VB, Beshnova DA, Vainshtein Y, Marth C, Mallm JP, Höfer T et al. (8 May 2014). "Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development". *Genome Research* 24: 1285–1295. doi:10.1101/gr.164418.113.
10. James H. Sang (2001-06-23). "Drosophila melanogaster: The Fruit Fly". In Eric C. R. Reeve. *Encyclopedia of genetics*. USA: Fitzroy Dearborn Publishers, I. p. 157. ISBN 978-1-884964-34-3. Retrieved 2009-07-01.
11. Adams MD, Celniker SE, Holt RA et al. (2000). "The genome sequence of Drosophila melanogaster". *Science* 287 (5461):2185–95. Bibcode:2000Sci...287.2185.. doi:10.1126/science.287.5461.2185. PMID 10731132. Retrieved 2007-05-25.
12. "NCBI (National Center for Biotechnology Information) Genome Database". Retrieved 2011-11-30.
13. Halligan DL, Keightley PD (2006). "Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison". *Genome Research* 16 (7): 875–84. doi:10.1101/gr.5022906. PMC 1484454. PMID 16751341.
14. Reiter, LT; Potocki, L; Chien, S; Gribskov, M; Bier, E (2001). "A Systematic Analysis of Human Disease-Associated Gene Sequences In Drosophila melanogaster". *Genome Research* 11 (6): 1114–1125. doi:10.1101/gr.169101. PMC 311089. PMID 11381037.
15. Raney BJ, Cline MS, Rosenbloom KR, Dreszer TR, Learned K, Barber GP, Meyer LR, Sloan CA, Malladi VS, Roskin KM, Suh BB, Hinrichs AS, Clawson H, Zweig AS, Kirkup V, Fujita PA, Rhead B, Smith KE, Pohl A, Kuhn RM, Karolchik D, Haussler D, Kent, WJ (January 2011).

- "ENCODE whole-genome data in the UCSC genome browser (2011 update)". *Nucleic Acids Res.* 39 (Database issue): D871–5. doi:10.1093/nar/gkq1017. PMC 3013645. PMID 21037257.
16. Maher B (September 2012). "ENCODE: The human encyclopaedia". *Nature* 489 (7414): 46–8. doi:10.1038/489046a. PMID 22962707.
 17. Saey, Tina Hesman (6 October 2012). "Team releases sequel to the human genome". Society for Science & the Public. Retrieved 18 October 2012.
 18. "The modENCODE Project: Model Organism ENCyclopediaOf DNA Elements (modENCODE)". NHGRI website. Retrieved 2008-11-13.
 19. Celniker S (2009-06-11). "Unlocking the secrets of the genome". *Nature*.
 20. Youngson, Robert M. (2006). *Collins Dictionary of Human Biology*. Glasgow: HarperCollins. ISBN 0-00-722134-7.
 21. Strahl BD, Allis CD (Jan 2000). "The language of covalent histone modifications". *Nature* 403 (6765): 41–5. doi:10.1038/47412. PMID 10638745.
 22. Ning Song, Jie Liu, Shucaian, Tomoya Nishino, Yoshitaka Hishikawa and Takehiko Koji (2011). "Immunohistochemical Analysis of Histone H3 Modifications in Germ Cells during Mouse Spermatogenesis". *Acta Histochemica et Cytochemica* 44 (4): 183–90. doi:10.1267/ahc.11027. PMC 3168764. PMID 21927517.
 23. Cao R, Wang L, Wang H, Xia L, Erdjument-Bromage H, Tempst P et al. (2002). "Role of histone H3 lysine 27 methylation in Polycomb-group silencing". *Science* 298 (5595): 1039–43. doi:10.1126/science.1076997. PMID 12351676.
 24. deNapoles M, Mermoud JE, Wakao R, Tang YA, Endoh M, Appanah R et al. (2004). "Polycomb group proteins Ring1A/B link ubiquitylation of histone H2A to heritable gene silencing and X inactivation". *Dev Cell* 7 (5): 663–76. doi:10.1016/j.devcel.2004.10.005. PMID 15525528.
 25. Krogan NJ, Dover J, Wood A, Schneider J, Heidt J, Boateng MA, Dean K, Ryan OW, Golshani A, Johnston M, Greenblatt JF, Shilatifard A (March 2003). "The Paf1 complex is required for histone H3 methylation by COMPASS and Dot1p: linking transcriptional elongation to histone methylation". *Mol. Cell* 11 (3): 721–9. doi:10.1016/S1097-2765(03)00091-1. PMID 12667454.
 26. Krogan NJ, Dover J, Khorrani S, Greenblatt JF, Schneider J, Johnston M, Shilatifard A (March 2002). "COMPASS, a histone H3 (Lysine 4) methyltransferase required for telomeric silencing of gene expression". *J. Biol. Chem.* 277 (13): 10753–5. doi:10.1074/jbc.C200023200. PMID 11805083.
 27. Wood A, Schneider J, Dover J, Johnston M, Shilatifard A (2005). "The Bur1/Bur2 complex is required for histone H2B monoubiquitination by Rad6/Bre1 and histone methylation by COMPASS". *Mol Cell* 20 (4): 589–99. doi:10.1016/j.molcel.2005.09.010. PMID 16307922.
 28. Robzyk K, Recht J, Osley MA (2000). "Rad6-dependent ubiquitination of histone H2B in yeast". *Science* 287 (5452): 501–4. doi:10.1126/science.287.5452.501. PMID 10642555.
 29. Wenxiu Ma, Wing Hung Wong, stanford university (2011). "Chapter Three – The Analysis of ChIP-Seq Data". *Methods in Enzymology Volume 497, 2011, Pages 51–73 Synthetic Biology, Part A*

30. H. Ji, H. Jiang, W. Ma, D.S. Johnson, R.M. Myers, W.H. Wong An integrated software system for analyzing ChIP-chip and ChIP-seq data *Nat. Biotechnol.*, 26 (2008), pp. 1293–1300
31. W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, D. Haussler The human genome browser at UCSC *Genome Res.*, 12 (2002), pp. 996–1006
32. Elizabeth G. Wilbanks, Marc T. Facciotti. Evaluation of Algorithm Performance in ChIP-Seq Peak Detection. Published: July 8, 2010 DOI: 10.1371/journal.pone.0011471
33. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (June 2002). "The human genome browser at UCSC". *Genome Res.* 12 (6): 996–1006. doi:10.1101/gr.229102. PMC 186604. PMID 12045153.
34. Bailey TL, Williams N, Misleh C, Li WW (2006). "MEME: discovering and analyzing DNA and protein sequence motifs". *Nucleic Acids Res* 34 (Web Server issue): W369–373. doi:10.1093/nar/gkl198. PMC 1538909. PMID 16845028.
35. Timothy L. Bailey, "DREME: Motif discovery in transcription factor ChIP-seq data", *Bioinformatics*, 27(12):1653-1659, 2011.
36. Principles of nucleation of H3K27 methylation during embryonic development. Simon J. van Heeringen, Robert C. Akkers, Ila van Kruijsbergen, M. Asif Arif, Lars L.P. Hanssen, NilofarSharifi, and Gert Jan C. Veenstra. *Genome Res.* 2014 Mar; 24(3): 401–410. doi: 10.1101/gr.159608.113, PMCID: PMC3941105
37. Epigenetic repression of the Igk locus by STAT5-mediated recruitment of the histone methyltransferase Ezh2 Malay Mandal, Sarah E Powers, Mark Maienschein-Cline,
38. Elizabeth T Bartom, Keith M Hamel, Barbara L Kee, Aaron R Dinner & Marcus R Clark. *Nature Immunology* 12, 1212–1220 (2011) doi:10.1038/ni.2136
39. Alberts, B.; Bray, D.; Lewis et al., J. (2002): *Molecular Biology of the Cell.* , ISBN: 0-8153-4072-9
40. NCBI map viewer, and A. B. Carvalho, 2002. *Curr. Op. Genet. & Devel.* 12:664-668
41. en.wikipedia.org/wiki/Histone

APPENDIX-1 (SCRIPTS)

1) peak.pl:

```
sub main()
{
    print "enter the file name with extension: ";
    $string = <STDIN>;
    print "\n\nreading and processing file...\n\n ";
    my @words = split / /, $string;
    $input = $words[0];
    if (scalar @words eq 4)
    {
        if (substr($words[1],0,2) eq "-s") ## chromosome regions
        {
            $words[1] =~ s/[^d.]//g;
            $chrRegionThreshold= $words[1];
        }
        else
        {
            print "\nparameters are not correctly defined. results may not be
appropriate... \n";
        }
        if (substr ($words[2],0,2) eq "-p") ## min length of the peak
        {
            $words[2] =~ s/[^d.]//g;
            $MinPeakLength = $words[2];
        }
        else
        {
            print "\nparameters are not correctly defined. results may not be
appropriate... \n";
        }
        if (substr ($words[3],0,2) eq "-h") ## height constrain for peak
        {
            $words[3] =~ s/[^d.]//g;
            $peakHeight = $words[3];
        }
        else
    }
}
```

```

        {
            print "\nparameters are not correctly defined. results may not be
appropriate... \n";
        }
    }
    else ## default values
    {
        $chrRegionThreshold= 0.2;
        $MinPeakLength = 4;
        $peakHeight = 0.3;
    }
    open ($fh, $input) or die "could not open file: $input due to error: $!";
    while ($row =<$fh>)
    {
        $firstElement = substr($row,0,1);
        chomp $row;
        if (index($firstElement, "#") == -1 and $row ne "")
        {
            push (@all, "$row");
        }
    }
    $aboutFile = shift @all;
    $parameter = shift @all;
    close $fh;

#####
## 1. File description extraction##
#####
if (index($parameter, "variableStep") != -1 and index($aboutFile, "track type=wiggle_0")
!= -1)
{
    @parts = split /(chrom=| )/, $parameter;
    $chromosome = $parts[4];
    @parts = split /(span=| )/, $parameter;
    $span = $parts[6];

    my ($stringency , $avg, $maxHeight, $rMean) = access_values(\@all);
    $avg = int $avg;
    foreach my $t (@all)
    ## isolation of values based on stringency

```

```

    {
        my @parts = split /(\t)/ , $t;
        if ($parts[2]> $stringency)
        {
            push (@peaks, $t);
            push (@no, $parts[0]);
        }
    }

#####
## 2. extraction of the peak points##
#####

if ($span eq "")
{
    $span = $avg;

    push (@allData, @peaks);
    push (@allData, @all);
    @allData = sort { $a <=> $b } @allData;
    for ( my $e = 0; $e< scalar @allData; $e++ )
    {
        if ($allData[$e] eq $allData[$e+1])
        {
            if ($allData[$e+2] eq $allData[$e+3])
            {
                push(@temp, "$allData[$e]");
            }
            else
            {
                push (@temp, "$allData[$e]");
                push (@a, [@temp]); ## pushing in 2D array ##
                @temp = ();
            }
        }
    }
}

}

else

```

```

{
element
    push (@no, 10000000000000000); ### if use this @no array then remove the last

    for (my $z =0; $z< scalar @no-1; $z++)
    ## labelling the peak NO.
    {
        if ($no[$z+1]-$no[$z] < $span + .1* $span)
        {
            push (@temp, "$peaks[$z]");
        }

        else
        {
            push (@temp, "$peaks[$z]");
            push (@a, [@temp]); ## pushing in the 2D array ##
            @temp = ();
        }
    }
}

#####
## 3. improving the peak selection##
#####

## peak width: limit: at least 4 values in the peak ##

foreach my $r (@a)
{
    if (scalar @{$r}> $MinPeakLength)
    {
        push (@as , $r);
    }
}

if (-e "results.wig")
{
    open my $file, '+<', "results.wig" or die "could not open the file: $out *error* $!";
    my $firstLine = <$file>;
    if (index($firstLine, "track type=wiggle_0") == -1)
    {
        print $file $aboutFile, "\n";
    }
}

```

```

    }
    close $file;
}
else
{
    open my $file, '>', "results.wig" or die "could not open the file: $out *error* $!";
    print $file $aboutFile, "\n";
    close $file;
}

## peak height: height must be more than the double of the mean otherwise no peak
present in the data

open (my $fileHandle, '>>', "results.wig") or die "could not open the file: $out *error* $!";
## opening the file for printing the results

if ($mean == 0.1 && $rMean == 0.1) #if the values are negative then to set the minimum
peak height to 1
{
    $rMean = 0.5;
}

if ($maxHeight < 2* $rMean) ## if no peak is present
{
    print $fileHandle $parameter, "\n";
    print $fileHandle "# no values to display: no peaks found \n";
    print "# no values to display: no peaks found \n";
}
else ## if any peak is present
{
    for ($r = 0; $r < scalar @as; $r++)
    {
        for my $elem (@{$as[$r]})
        {
            my ($no, $value) = split " ", $elem;
            if ($value > $maxHeight*$peakHeight and $value > 2*$rMean)
            {
                push (@peakArrayIndexAll , $r); ## index of the peak
array that are actual peaks (satisfying the above condition)
            }
        }
    }
}

```

```

    }
  }
}

## removing the repeated index values
foreach my $var ( @peakArrayIndexAll )
{
  if ( ! grep( /$var/, @peakArrayIndex ) )
  {
    push( @peakArrayIndex, $var );
  }
}

## extracting the array on the basis of index values
for my $rt ( @peakArrayIndex )
{
  push ( @realPeak, $as[$rt]);
}
print "\n no of peaks in this chromosome are: ", scalar @realPeak, "\n\n";
##printing the result
print $fileHandle $parameter, "\n";
for $rt(@realPeak)
{
  for $er(@{$rt})
  {
    print $fileHandle $er, "\n";
  }
}
close $fileHandle;

#####
## 4. extracting the chromosome no. from each peak ##
#####
open ($fh2, '>>', "start_point.txt") or die "can't open the file *error* $!";
open ($fh3, '>>', "end_point.txt") or die "can't open the file *error* $!";
open ($fh4, '>>', "peak_regions.txt") or die "can't open the file *error* $!";
for my $u ( @realPeak )
{
  @first = split /\t/, ${$u}[0];
  @last = split /\t/, ${$u}[-1];
}

```

```

    if ($first[0]<500)
    {
        print $fh2 $chromosome,"t","0","t",$first[0],"n";
    }
    else
    {
        print $fh2 $chromosome,"t",$first[0]-500,"t",$first[0],"n";
    }

    print $fh3 $chromosome,"t",$last[0],"t",$last[0]+500,"n";
    print $fh4 $chromosome,"t",$first[0],"t",$last[0],"n";
}
close $fh2;
close $fh3;
close $fh4;

#####
# 5. writing a log file #
#####

open ($fh5, '>>', "analysis.log") or die "can't open the file %error in log file% $!";
print $fh5 "*****\n
           this is a log file generate for chromosome: ",$chromosome,"n",
           "*****\n\n";

print $fh5 "span value: ",$span,"n";
print $fh5 "max peak height: ",$maxHeight,"n",
           "parameters:\n",-s (% region to be contained above mean):
",$chrRegionThreshold,"n",-p ( minimum length for the peak): ",$MinPeakLength,"n",
"-h ( minimum percent of the peak height relative to the heighest peak): ",$peakHeight,"n\n";
print $fh5 "stringency for the peak selection: ",$stringency,"n",
           "actual mean value for this chromosome: ",$rMean,"n\n";
print $fh5 "no of the peaks in the chrommosome: ",scalar @realPeak,"n\n";
close $fh5;

}

else
{

```



```

    print "this is not a wig file format.. or the file is not properly formatted";
}

}

#####
## accessing values ##
#####

subaccess_values()
{
    my @allValue = @{$_[0]};

    my $totalValue = 0;
    foreach my $x (@all)
    {
        my @break = split /\(t)/, $x;
        push (@chrNo, "$break[0]"); #contain the chromosome no.
        push (@chrValue, "$break[2]"); #contain the chromosome's signal value
        $totalValue = $totalValue+$break[2];
    }
    $mean = $totalValue/ scalar @chrValue;
    if ($mean < 0)
    {
        $mean = .1;
        $rMean = .1;
    }

    $sum = set_stringency();
    while ($sum/scalar @chrValue > $chrRegionThreshold)
    {
        $mean = $mean+$mean*.001;
        $sum = set_stringency();
    }

    print "mean = ".$mean."\n";

    ## set the span value if the span is not given
    if ($span eq "")

```

```

    {
        $avg = $chrNo[-1]/ scalar @chrNo;
    }

my $max = (sort { $b <=> $a } @chrValue)[0];
return ($mean, $avg, $max, $rMean);

}

#####
## dynamically sets the stringency for the given dataset ##
#####

subset_stringency()
{
    my $sum = 0;
    for (my $i = 0;$i< scalar @chrValue; $i++)
    {
        if ($chrValue[$i]>$mean)
        {
            $sum = $sum+1;
        }
    }
    return $sum;
}

main();

```

2) fastaconversion.pl:

```

use List::MoreUtilsqw(uniq);

print "enter the file name containing the chromosome co-ordinate (in UCSC format, read
readme.txt for more detail ):";
$input = <STDIN>;
open ($fh, $input) or die "could not open file: $input due to error: $!";

## making the array of the coordinates for each chromosome ##

```

```
#####
```

```
while ($row =<$fh>)  
{  
    @part = split /(\t)/, $row;  
    push (@all, $row);  
    push (@chrn, $part[0]);  
    push (@coordinates, $part[2]);  
    push (@end_coordinates, $part[4]);  
}
```

```
for (my $z =0; $z< scalar @chrn; $z++)  
{  
    if ($chrn[$z] eq $chrn[$z+1])  
    {  
        $tempo = $coordinates[$z].".".$end_coordinates[$z];  
        push (@temp,$tempo);  
    }  
  
    else  
    {  
        $tempo = $coordinates[$z].".".$end_coordinates[$z];  
        push (@temp, $tempo);  
        unshift (@temp, $chrn[$z]);  
        push (@chrPos, [@temp]); ## pushing the coordinates in the 2D array ##  
        @temp = ();  
    }  
}
```

```
## print the fasta format to the output file ##
```

```
#####
```

```
open($fh1, '>', "output.fa") or die "cant open the output file : $!";
```

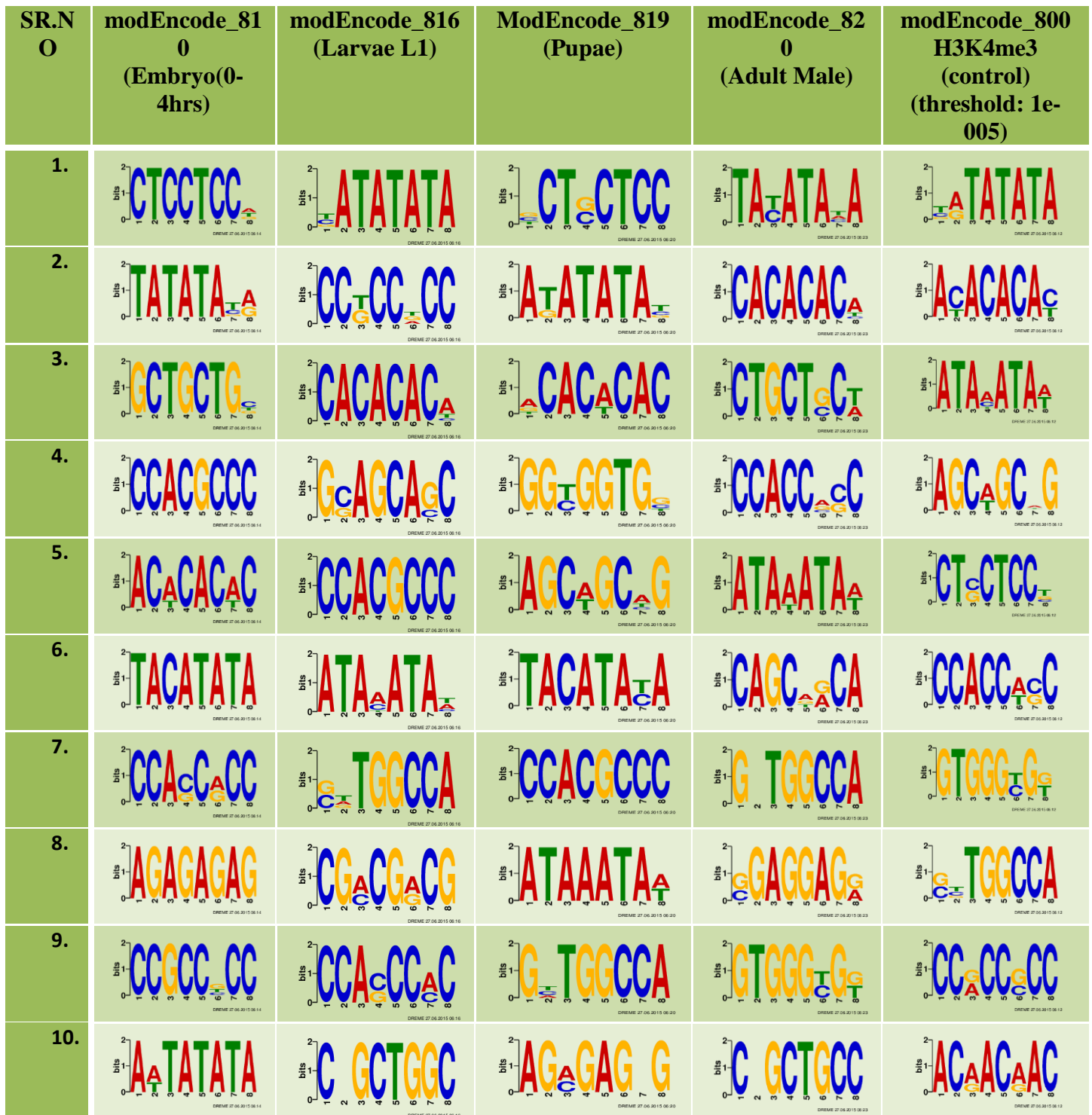
```
for $chrom (@chrPos)  
{  
    my $chromosome = shift @{$chrom};  
    $chromosomeFile = $chromosome.".fa";  
    open ($fileHandle,<', $chromosomeFile) or die "can't open the chromosome genome file  
: $!";  
    my $var = do { local $/; <$fileHandle> };
```

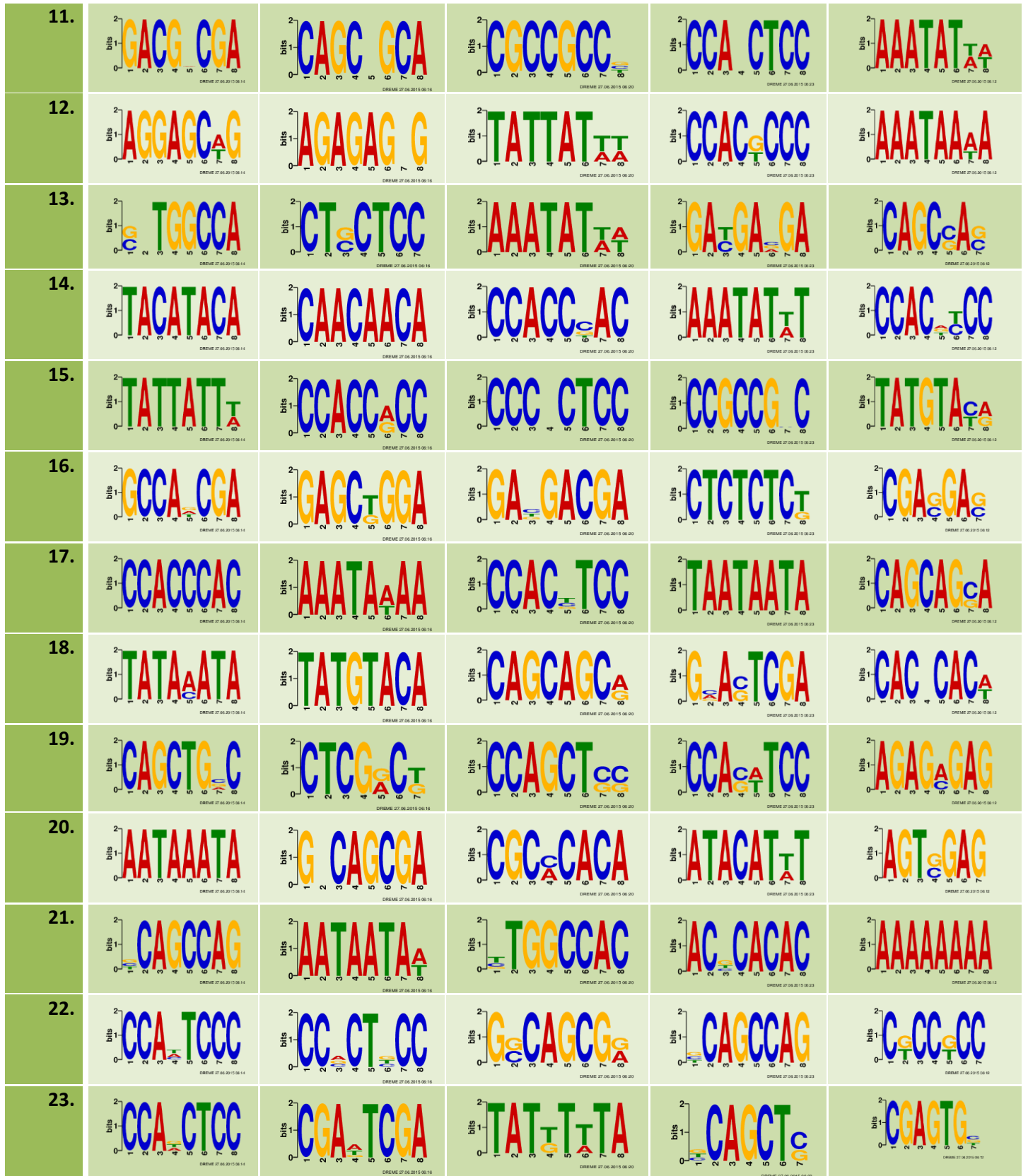
```

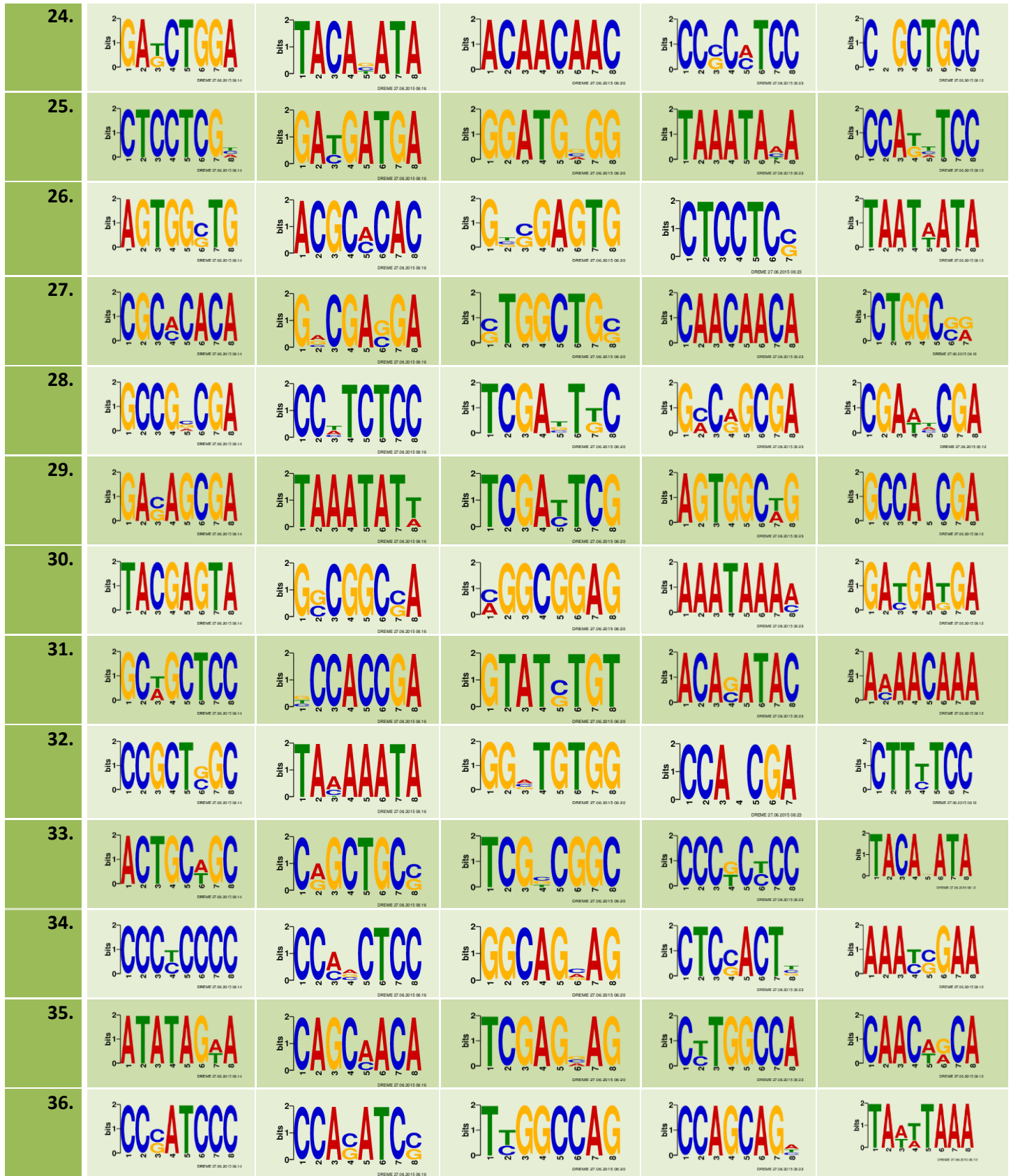
$var =~ s/>//;
$var =~ s/$chromosome//;
for (my $x =0; $x<scalar @{$chrom};$x++)
{
    @parts = split /(:/, @{$chrom}[$x];
    print $fh1 ">",$chromosome,":sig_start-", $x+1,":position:",$parts[0],"-
",$parts[2],"\n";
    $length = $parts[2]-$parts[0];
    print $fh1 substr $var,$parts[0],$length;
    print $fh1 "\n";
}
print "length of the $chromosome Chromosome: ",length($var),"\n";
}

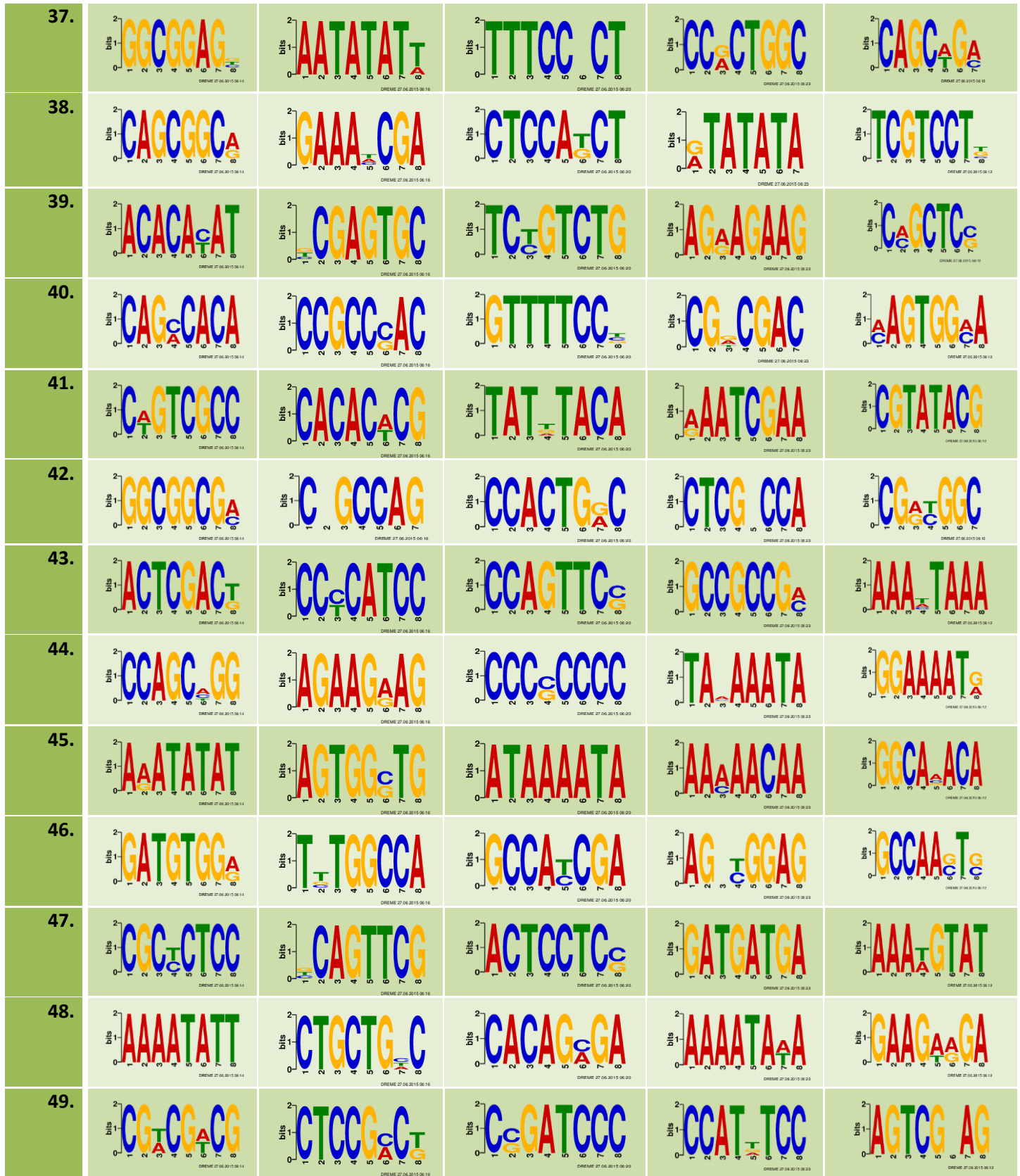
```

APPENDIX-2







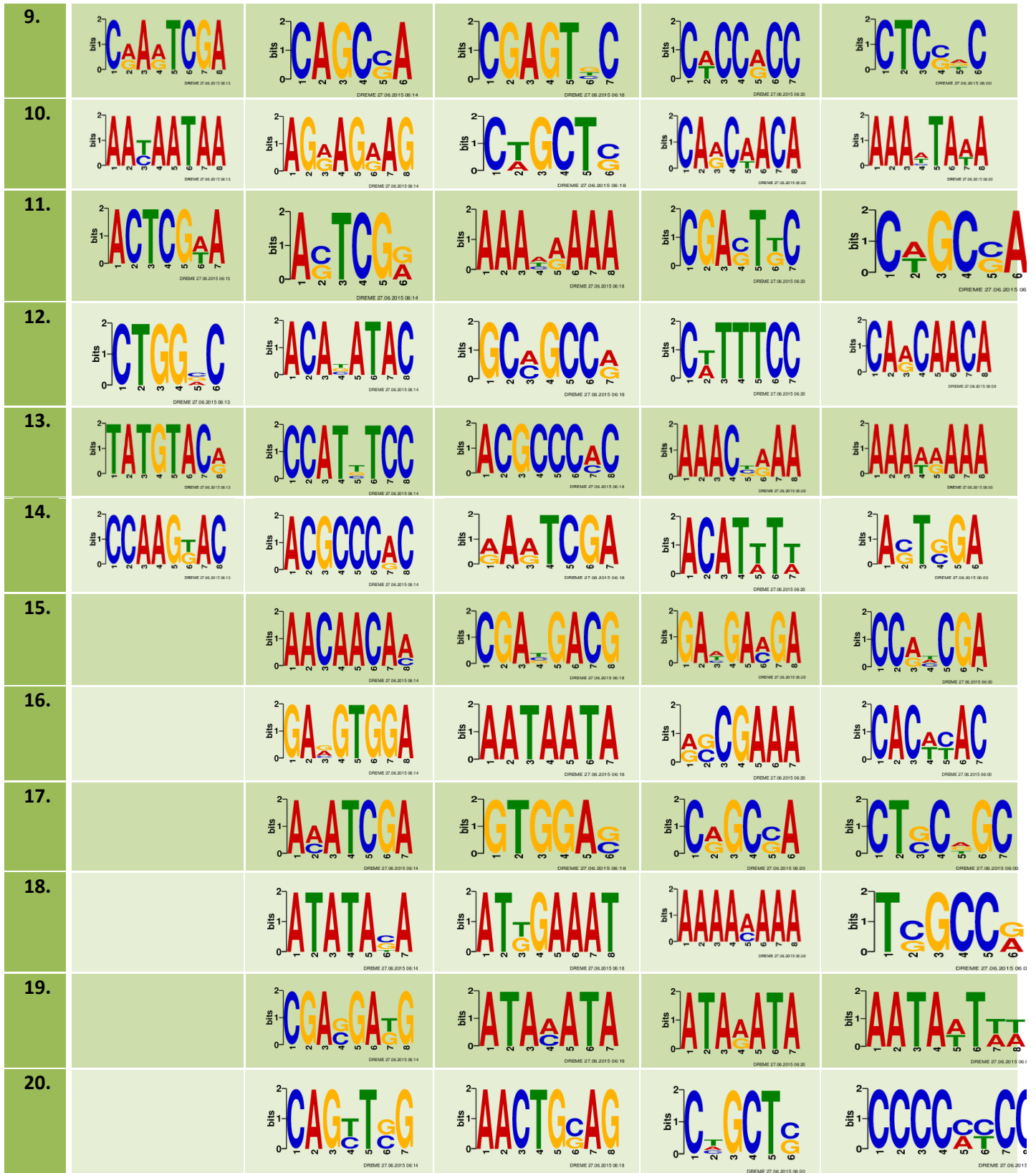


50.

















Table.5. Motifs present in the peak region of different dataset

SR. NO	modEncode_810 (Embryo(0-4hrs))	modEncode_816 (Larvae L1)	ModEncode_819 (Pupae)	modEncode_820 (Adult Male)	modEncode_800 H3K4me3 (control) (threshold: 1e-005)
1.					
2.					
3.					
4.					
5.					
6.					
7.					
8.					



21.				
22.				
23.				
24.				
25.				
26.				
27.				
28.				
29.				
30.				
31.				
32.				

33.					
34.					
35.					
36.					
37.					
38.					
39.					
40.					
41.					
42.					
43.					
44.					
45.					
46.					

47.					
48.					
49.					
50.					

Table.6. Motifs present in the end region of different dataset